Rewarding Coreference Resolvers for Being Consistent with World Knowledge

Aralikatte, Rahul; Lent, Heather; Gonzalez, Ana Valeria; Herschcovich, Daniel; Qiu, Chen; Sandholm, Anders; Ringaard, Michael; Søgaard, Anders

Published in:
Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

DOI:
10.18653/v1/D19-1118

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Aralikatte, R., Lent, H., Gonzalez, A. V., Herschcovich, D., Qiu, C., Sandholm, A., ... Søgaard, A. (2019). Rewarding Coreference Resolvers for Being Consistent with World Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1229-1235). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1118

Rewarding Coreference Resolvers for Being Consistent with World Knowledge

Rahul Aralikatte[1], **Heather Lent**[1], **Ana Valeria Gonzalez**[1], **Daniel Hershcovich**[1], **Chen Qiu**[1,2], **Anders Sandholm**[3], **Michael Ringaard**[3], and **Anders Søgaard**[1,3]

[1]University of Copenhagen, [2]China University of Geosciences, [3]Google Research
{rahul,hcl,ana,hershcovich,chen.qiu,soegaard}@di.ku.dk,
{sandholm,ringgaard}@google.com

Abstract

Unresolved coreference is a bottleneck for relation extraction, and high-quality coreference resolvers may produce an output that makes it a lot easier to extract knowledge triples. We show how to improve coreference resolvers by forwarding their input to a relation extraction system and reward the resolvers for producing triples that are found in knowledge bases. Since relation extraction systems can rely on different forms of supervision and be biased in different ways, we obtain the best performance, improving over the state of the art, using multi-task reinforcement learning.

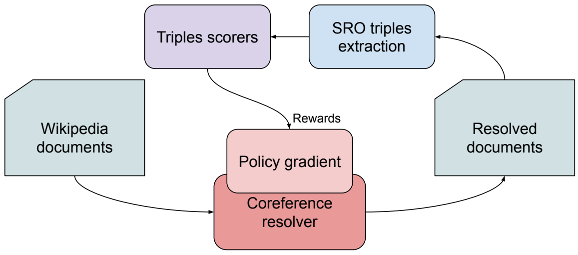

Figure 1: Our strategy for training a coreference resolver using reward from relation extraction.

1 Introduction

Coreference annotations are costly and difficult to obtain, since trained annotators with sufficient world knowledge are necessary for reliable annotations. This paper presents a way to *simulate* annotators using reinforcement learning. To motivate our approach, we rely on the following example from Martschat and Strube (2014, colors added to mark entity mentions):

(1) [**Lynyrd Skynyrd**]$_1$ was formed in **Florida**$_2$. Other bands from [**the Sunshine State**]$_2$ include **Fireflight** and **Marilyn Manson**.

Martschat and Strube (2014) cite the association between Florida and the Sunshine State as an example of a common source of name-name recall error for state-of-the-art coreference resolution systems. The challenge is that the two names co-occur relatively infrequently and are unlikely to do so in a moderate-sized, manually annotated training corpus. A state-of-the-art system may be able to infer the relation using distributional information about the phrase the Sunshine State, but is likely to have limited evidence for the decision that it is coreferential with Florida rather than Lynyrd Skynyrd.

While coreference-annotated data is scarce, knowledge bases including factual information (such as that Fireflight is from Florida) are increasingly available. For a human annotator unaware that Florida is sometimes referred to as the Sunshine State, the information that Fireflight is from Florida is sufficient to establish that Florida and the Sunshine State are (with high probability) coreferential. This paper explores a novel architecture for making use of such information from knowledge bases by tying a coreference resolution system to a relation extraction system, enabling us to reward the coreference system for making predictions that lead us to infer facts that are consistent with such knowledge bases. This potentially provides us with more evidence for resolving coreference such as (1).

We propose a training strategy (Figure 1) in which we pass on the predictions of a neural coreference resolver to an open relation extraction (OpenRE) system, matching relations extracted from resolved sentences with a knowledge base. We show how checking the produced relationships for consistency against the knowledge base produces a reward that is, indirectly, a signal about the quality of the coreference resolution. In order to generalize this signal beyond the coverage of the knowledge base, we train a Universal Schema model (Riedel et al., 2013) and use its confidence as our reward function. With this reward function,

	RE-KG	RE-Text	RE-Joint
Document	NA	Bach was a German composer. He is known for instrumental compositions such as the Art of Fugue	Bach was a German composer. He is known for instrumental compositions such as the Art of Fugue
OpenRE	NA	(Bach, born in, Germany) (Bach, occupation, Composer) (He, composition, Art of Fugue)	(Bach, born in, Germany) (Bach, occupation, Composer) (He, composition, Art of Fugue)
Verification	NA	NA	✗ (Bach, born in, Germany) ✓ (Bach, occupation, Composer) ✗ (He, composition, Art of Fugue)
Result	(Bach, born in, Eisenach) (Bach, born on, 21 March 1685) (Bach, occupation, Composer) ...	(Bach, born in, Germany) (Bach, occupation, Composer) (He, composition, Art of Fugue)	(Bach, occupation, Composer)

Figure 2: The columns show the different pipelines used to obtain data for training the reward models. The pipeline for: (i) RE-KG directly extracts triples from Wikidata, (ii) RE-Text runs Wikipedia summaries through OpenRE to generate triples, and (iii) RE-Joint adds an additional verification step by checking if the generated triples exist in Wikidata.

we do policy-gradient fine-tuning of our coreference resolver, effectively optimizing its predictions' consistency with world knowledge.

Contributions We demonstrate that training a coreference resolver by reinforcement learning with rewards from a relation extraction system, results in improvements for coreference resolution. Our code is made publicly available at `https://github.com/rahular/coref-rl`

2 Consistency Reward for Coreference Resolution

In order to reward a coreference resolver for being consistent with world knowledge, we propose a simple training strategy based on relation extraction: (i) Sample a Wikipedia[1] document at random, (ii) Replace mentions with their antecedents using a coreference resolver, (iii) Apply an off-the-shelf openRE system to each rewritten document, (iv) Score relationships that include coreferent mentions using Universal Schema, and (v) Use the score as a reward for training the coreference resolvers.

Reward functions To model consistency with world knowledge, we train different Universal Schema models (Riedel et al., 2013; Verga and McCallum, 2016), resulting in three reward functions (Figure 2): **RE-KG** (Knowledge Graph Universal Schema) is trained to predict whether two entities are linked in Wikidata[2]; **RE-Text** (Text-based Universal Schema) is trained to predict whether two entities co-occur in Wikipedia; and **RE-Joint** (Joint Universal Schema) is trained to predict whether two entities are linked *and* co-occur. The three rewards focus on different aspects of relationships between entities, giving complimentary views of what entities are related.

Similar to Verga et al. (2016), we parameterize candidate relation phrases with a BiLSTM (Graves and Schmidhuber, 2005), and use pre-trained Wikidata BigGraph embeddings (Lerer et al., 2019) as the entity representations. We apply a one-layer MLP on the concatenated representations to get the reward value.

Updating the coreference resolver Each resolved document is converted into n subject-relation-object (SRO) triples by an open information retrieval system (Angeli et al., 2015). Each triple t_i is then scored using a reward function to obtain a reward r_i for $i \in \{1, \ldots, n\}$. The final document-level reward is the normalized sum of the individual rewards as shown in Equation 1, where R_h is a moving window containing the previous $h = 100$ normalized reward values.

$$R = \frac{\sum_i r_i - mean(R_h)}{stddev(R_h)} \tag{1}$$

Since R is not differentiable with respect to the coreference resolver's parameters, we use pol-

[1] `https://www.wikipedia.org`

[2] `https://www.wikidata.org`

icy gradient training to update the coreference resolver. We select the best action according to the current policy, using random exploration of the alternative solutions with $p = \frac{1}{10}$.

Multi-task reinforcement learning Our overall training procedure is presented in Algorithm 1. After training the three aforementioned reward models, we create **RE-Distill** by interpolating their trained weights. Next, we pre-train a coreference resolver using supervised learning, and fine-tune it using each of the three reward functions to get three different coreference policies: **Coref-KG**, **Coref-Text** and **Coref-Joint**, respectively. We then use multi-task reinforcement learning to combine these three policies to get **Coref-Distill**. Our approach is a particular instance of DisTraL (Teh et al., 2017), using policy gradient and model interpolation. Finally, **Coref-Distill** is fine-tuned with rewards from **RE-Distill**.

Algorithm 1 Multi-task Reinforcement Learning

Require: Baseline initialized policies θ_n for $n \in \{1, 2, 3\}$
Require: Reward functions $\texttt{reward}_n$ for $n \in \{1, 2, 3\}$
Require: Distilled reward function $\texttt{reward}_*$
  while stopping criterion not met **do**
    Sample k documents D^k
    for $d \in D^k$ **do**
      for $n \in \{1, 2, 3\}$ **do**
        $\mathcal{C}_d$ = entity clusters with θ_n
        d' = resolve d with $\mathcal{C}_d$
        $\mathcal{T}$ = obtain OpenIE triples for d'
        $r = \text{reward}_n(d')$
        $\hat{g}_k$ = policy gradient for θ_n with reward r
        $\theta_n^{k+1} = \theta_n^k + \alpha_k \hat{g}_k$
      end for
    end for
  end while
  Distilled policy $\theta_* = \frac{\theta_1 + \theta_2 + \theta_3}{3}$
  Sample k documents D^k
  for $d \in D^k$ **do**
    d' = resolve d with $\mathcal{C}_d$
    $\mathcal{T}$ = obtain OpenIE triples for d'
    $r = \text{reward}_*(d')$
    $\hat{g}_k$ = policy gradient for θ_* with reward r
    $\theta_*^{k+1} = \theta_*^k + \alpha_k \hat{g}_k$
  end for
  return Distilled policy θ_*

3 Experiments

We use a state-of-the-art neural coreference resolution model (Lee et al., 2018) as our baseline coreference resolver.[3] This model extends Lee et al. (2017) with coarse-to-fine inference and ELMo pretrained embeddings (Peters et al., 2018).

[3] https://github.com/kentonl/e2e-coref

System	Data	Accuracy	F_1 score
RE-KG	12M	0.64	0.78
RE-Text	2M	0.71	0.83
RE-Joint	60K	0.58	0.73
RE-Distill	—	**0.78**	**0.88**

Table 1: Training data size, accuracy and F_1 scores of the reward models on the 200,000 validation triples.

Data We use the standard training, validation, and test splits from the English OntoNotes.[4] We also evaluate on the English WikiCoref (Ghaddar and Langlais, 2016), with a validation and test split of 10 and 20 documents respectively.

Reward model training We use data from English Wikipedia and Wikidata to train our three reward models. For training **RE-KG**, we sample 1 million Wikidata triples, and expand them to 12 million triples by replacing relation phrases with their aliases. For **RE-Text**, we pass the summary paragraphs from 50,000 random Wikipedia pages to Stanford's OpenIE extractor (Manning et al., 2014), creating 2 million triples. For **RE-Joint**, we only use Wikipedia triples that are grounded in Wikidata, resulting in 60,000 triples.[5] We further sample 200,000 triples from Wikidata and Wikipedia for validation, and train the reward models with early stopping based on the F_1 score of their predictions.

Evaluation All models are evaluated using the standard CoNLL metric, which is the average F_1 score over MUC, CEAFe, and B^3 (Denis and Baldridge, 2009).

4 Results

Since the quality of our reward models is essential to the performance of the coreference resolver adaptations, we first report the validation accuracy and F_1 scores of the four reward models used, in Table 1. We clearly see the advantage of distillation, with a 5% absolute difference between the best single model (**RE-Text**) and **RE-Distill**.

Table 2 presents the downstream effects of applying these reward functions to our baseline

[4] https://catalog.ldc.upenn.edu/LDC2013T19

[5] That is, we retain only those triples whose subject and object can be linked to an entity in Wikidata.

System	OntoNotes	WikiCoref
Lee et al. (2018)	72.60	57.49
Coref-KG	72.96	57.84
Coref-Text	72.99	57.54
Coref-Joint	72.77	57.51
Coref-Distill	**73.10**	**58.14**

Table 2: Coreference results: average F_1 scores on the OntoNotes and WikiCoref test sets. Differences are significant w.r.t. B^3 (bootstrap test, $p < 0.05$).

coreference policy.[6]

The coreference resolution results are similar to the relation extraction results: using a distilled policy, learned through multi-task reinforcement learning, leads to better results on both datasets.[7]

While improvements over the current state of the art are relatively small, they reflect significant progress, as they demonstrate the ability to successfully augment coreference resolvers with "free" data from large-scale KB like Wikidata. For relation extraction, this could have positive downstream effects, and also ensure that relations are consistent with real world knowledge. Moreover, this approach has the potential to also be beneficial for coreference resolution in low resource languages, where less annotated data is available, as Wikidata triples are abundant for many languages.

5 Analysis

Empirically, we find that fine-tuning the coreference resolver on Wikidata results in two kinds of improvements:

Better mention detection Since the model is rewarded if the SRO triples produced from the resolved document are present in Wikidata, the model can do well only if it correctly resolves the subject and object, which are usually named entities (more generally, noun phrases). Indeed, we see an improvement in mention detection as exemplified in the first example of Figure 3. Compared to the baseline, the fine-tuned model identifies a larger number of entities, including "southern hemisphere", "Cambridge" and "Oxford", which are missed by the baseline model.

[6]The models were re-trained from scratch, and the scores are slightly different from those reported in Lee et al. (2018).

[7]We repeat this experiment three times with different random seeds and observed the same pattern and very robust performance across the board.

Better linking As a direct consequence of the above, the model is inclined to also link noun phrases that are not entities. In the second example of Figure 3, we see that "This attempt" is linked to "releasing" by the fine-tuned model. Interestingly, we do not see this type of *eventive* noun phrase linking either in OntoNotes or in the predictions of the baseline model.

This phenomenon, however, also has a side-effect of producing singleton clusters and spurious linking, which adversely affect the recall. On the OntoNotes test data, while the average precision of the best performing fine-tuned model is higher than the baseline (75.62 vs. 73.80), a drop in recall (70.75 vs. 71.34) causes the final F_1 score to only marginally improve.

6 Related Work

Coreference resolution Among neural coreference resolvers (Wu and Ma, 2017; Meng and Rumshisky, 2018), Lee et al. (2017) were the first to propose an end-to-end resolver which did not rely on hand-crafted rules or a syntactic parser. Extending this work, Lee et al. (2018) introduced a novel attention mechanism for iteratively ranking spans of candidate coreferent mentions, thereby improving the identification of long distance coreference chains. Zhang et al. (2019) improve pronoun coreference resolution by 2.2 F1 using linguistic features (gender, animacy and plurality) and a frequency based predicate-argument selection preference as external knowledge. Emami et al. (2018) incorporate knowledge into coreference resolution by means of information retrieval, finding sentences that are syntactically similar to a given instance, and improving F1 by 0.16.

Reinforcement learning RL has been used for many NLP tasks, including coreference resolution (Clark and Manning, 2016) and relation extraction (Zeng et al., 2018). Clark and Manning (2016) use RL to improve coreference resolution by optimizing their mention ranking model and directly use the standard evaluation metrics as the rewards. We, on the other hand, perform end-to-end optimization by rewarding the model's consistency with real world knowledge using relation extraction. To our knowledge, we are the first to use consistency with world knowledge as a reward for

	Baseline system	Fine-tuned system
Mention detection	According to the library's publications, it is the largest academic library in the southern hemisphere. The university has a number of residential college and halls of residence, based on the college system of Cambridge and Oxford universities.	According to the library's publications, it is the largest academic library in the southern hemisphere. The university has a number of residential college and halls of residence, based on the college system of Cambridge and Oxford universities.
Linking	On March 19, Obama continued his outreach to the Muslim world, releasing a New Year's video message to the people and government of Iran. This attempt was rebuffed by the Iranian leadership.	On March 19, Obama continued his outreach to the Muslim world, releasing a New Year's video message to the people and government of Iran. This attempt was rebuffed by the Iranian leadership.

Figure 3: Mention detection and linking examples by the baseline system from Lee et al. (2018), and the best performing fine-tuned system (Coref-Distill). Mentions of the same color are linked to form a coreference cluster.

tasks other than knowledge base construction.[8]

Knowledge bases Knowledge bases have been leveraged across multiple tasks across NLP (Bordes et al., 2011; Chang et al., 2014; Lin et al., 2015; Toutanova et al., 2015; Yang and Mitchell, 2017). Specifically for coreference resolution, Prokofyev et al. (2015) implement a resolver that ensures semantic relatedness of resulting coreference clusters by leveraging Semantic Web annotations. Their work incorporates knowledge graph information only in the final stage of the resolver's pipeline, and not during training. In contrast, our work augments information from the knowledge base directly into the training pipeline. Also, they use DBpedia (Auer et al., 2007) as the ontology. Although both Wikidata and DBpedia are designed to support working with Wikipedia articles, DBpedia can be considered as a subset of Wikidata as Wikipedia infoboxes are its main data source. The advantage of Wikidata over DBpedia is its size, and the fact that it is multilingual, which will allow applying our method to other languages in the future.

7 Conclusion

We presented an architecture for adapting coreference resolvers by rewarding them for being consistent with world knowledge. Using simple multi-task reinforcement learning and a knowledge extraction pipeline, we achieved improvements over the state of the art across two datasets. We believe this is an important first step in exploring the usefulness of knowledge bases in the context of coreference resolution and other discourse-level phenomena. In this area, manually annotated data is particularly expensive, and we believe leveraging knowledge bases will eventually reduce the need for manual annotation.

[8] Mao et al. (2018), for example, use reinforcement learning with consistency-like reward to induce lexical taxonomies.

Acknowlegments

We thank the reviewers for their valuable comments. Rahul Aralikatte, Daniel Hershcovich, Heather Lent, and Anders Søgaard are funded by a Google Focused Research Award. Heather Lent is also funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801199. Chen Qiu is funded in part by the National Natural Science Foundation of China under grant No. 61773355 and the China Scholarship Council.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Seman-*

tic Web Conference, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *AAAI*.

Kai-Wei Chang, Wen tau Yih, Bishan Yang, and Christopher Meek. 2014. Typed tensor decomposition of knowledge bases for relation extraction. In *EMNLP*.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42.

Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A generalized knowledge hunting framework for the Winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 25–31, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA), European Language Resources Association (ELRA).

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. PyTorch-BigGraph: A Large-scale Graph Embedding System. In *Proceedings of the 2nd SysML Conference*, Palo Alto, CA, USA.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2015. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Yuning Mao, Xiang Ren, Jiaming Shen, Xiaotao Gu, and Jiawei Han. 2018. Building a large-scale annotated chinese corpus. In *ACL*.

Sebastian Martschat and Michael Strube. 2014. Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2070–2081, Doha, Qatar. Association for Computational Linguistics.

Yuanliang Meng and Anna Rumshisky. 2018. Triad-based neural network for coreference resolution. In *COLING*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Roman Prokofyev, Alberto Tonon, Michael Luggen, Loic Vouilloz, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2015. Sanaphor: Ontology-based coreference resolution. In *International Semantic Web Conference*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia. Association for Computational Linguistics.

Yee Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. 2017. Distral: Robust multitask reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4496–4506. Curran Associates, Inc.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, San Diego, California. Association for Computational Linguistics.

Patrick Verga and Andrew McCallum. 2016. Row-less universal schema. In *AKBC*.

Jheng-Long Wu and Wei-Yun Ma. 2017. A deep learning framework for coreference resolution based on convolutional neural network. *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 61–64.

Bishan Yang and Tom Michael Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *ACL*.

Xiangrong Zeng, Shizhu He, Kang Liu, and Jian Zhao. 2018. Large scaled relation extraction with reinforcement learning. In *AAAI*.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 872–881, Minneapolis, Minnesota. Association for Computational Linguistics.